FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

INFERENCIA ESTATÍSTICA CON DATOS TRUNCADOS

Iria Portela Comba

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

INFERENCIA ESTATÍSTICA CON DATOS TRUNCADOS

Iria Portela Comba

Xullo 2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Inferencia Estatística con Datos Truncados
Breve descrición do contido
Os datos truncados xorden cando algúns individuos non son observables debido ao valor que toma a variable de interés ou outras variables relacionadas con ela. En concreto, este fenómeno preséntase nas observacións astronómicas, onde certas condicións de posición e luminosidade non son observables. Tamén xorden nas Análises de Supervivencia, onde por exemplo os individuos que se curan antes de ir ao médico, non chegan a ser observados. Para que os estimadores non estén sesgados, diseñaronse estimadores específicos con este tipo de datos. Están basados na suposición de independencia entre a variable de interés e as variables que truncan o proceso de observación. No caso da duración dunha enfermidade, sería a independencia entre o tempo de duración da enfermidade e o tempo que tarda o paciente en acudir ao médico. Neste traballo revisaranse os estimadores con datos truncados e ilustraranse con datos simuladas e datos reais.

Índice xeral

Resumo

Este traballo está centrado nos métodos estatísticos para a análise de datos truncados. Comezouse introducindo a relación que existe entre a distribución orixinal e a distribución truncada dunha poboación, creando unha serie de estimadores que permiten estimar características da poboación total a partir dunha mostra observable. Fíxose un estudo destes estimadores, indicando as súas propiedades principais. Ao longo de todo o traballo foise ilustrando con exemplos cada método que se presentaba, aparecendo, nun segundo capítulo, simulacións feitas en R coas súas respectivas análises. Ao final do traballo podemos atoparnos cun apéndice no cal se introducen conceptos básicos de supervivencia e as liñas de código de R que foron usadas nas simulacións.

Abstract

This project is focused on the statistical methods for the analysis of truncated data. It starts by introducing the relation that exists between the original distribution and the truncated distribution of a population, creating a series of estimators which allow estimating characteristics of the total population from an observable sample. We made a study of these estimators, indicating their main properties. Throughout the project, each method that was presented was illustrated with examples, appearing, in a second chapter, simulations made in R with their respective analyzes. At the end of the work we can find an appendix in which basic survival concepts and the lines of R code that were used in the simulations are introduced.

Introdución

O problema de truncamento aparece en diversas disciplinas. Destacamos como unha das que primeiro estudaron este problema a Astronomía.

A luminosidade que produce un obxecto astronómico defínese como o seu brillo a unha distancia fixa, como se observa dende a Terra. Moitas veces na vida nocturna só somos capaces de percibir certas estrelas, debido á distancia que presentan estas do noso planeta. Certos estudos analizaron o problema de truncamento e foron capaces de construír os posibles astros que hai e non podemos observar en certa parte do ceo. Algúns destes estudos están recollidos no artigo [4] do astrónomo Lynden-Bell.

Para isto créanse estimadores que permiten, a partir dos datos observados, obter unha nova construcción da mostra.

Noutras aplicacións, matemáticos usaron este tipo de modelo para aplicalo en estudos para analizar onde atopar reservas petrolíferas ou optimizar problemas de frenado, etc.

Con frecuencia podemos atoparnos estes casos en análises de supervivencia, que son un conxunto de técnicas empregadas para analizar o tempo que transcurre dende o punto inicial ata certo punto final dun suceso. Por exemplo, no artigo [2] estúdase o fenómeno de truncamento por retraso de notificación nos rexistros de SIDA.

Unha característica principal neste estudo é a fiabilidade, que é a probabilidade de que o suxeito cumpra determinada función en certo período de tempo. Para isto, falaremos do concepto de función de risco, que é a que se encarga de estimar esta probabilidade.

Neste campo da Estatística aparecen distintos conceptos os cales enlazan uns con outros, como poden ser a relación da supervivencia coa función de distribución e coa función de risco, as cales se poden expresar en termos de supervivencia tanto no caso discreto como no continuo. Todas estas relacións podémolas atopar no apéndice deste traballo.

Para facer todo estes estudos, consideraremos una poboación de tamaño N. Supoñamos as variables X e Y independentes e positivas, sendo os pares $(X_1, Y_1), ..., (X_N, Y_N)$ con $i \in \{1,...,N\}$ independentes entre sí, con distribucións F e G, respectivamente. Observaremos só aqueles (X_i, Y_i) que verifiquen $Y_i \leq X_i$, logo quedarémonos así con n observacións.

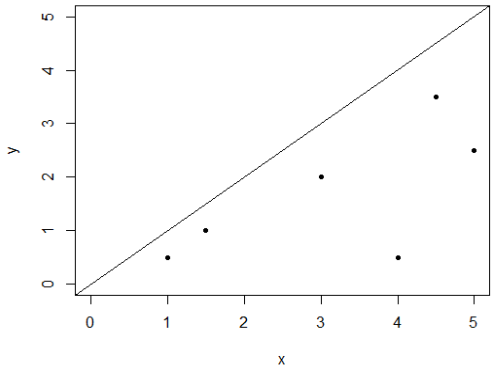
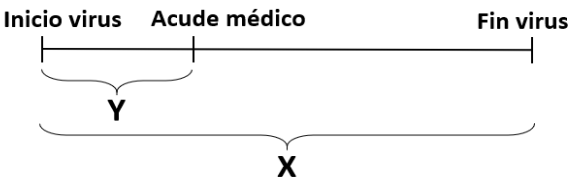


Figura 1: Observacións que verifican a condición $Y_i \leq X_i$

Falar de truncamento é falar de algo que está incompleto, e que so temos certos datos observables, como amosa a Figura 1, na cal só se observan os datos por debaixo da diagonal a partir dos cales debemos construír o total da poboación.

Existen dous tipos de truncamento: o truncamento pola esquerda e o truncamento pola dereita. Na nosa condición de truncamento $Y \leq X$, X está truncada pola esquerda por Y e Y está truncada pola dereita por X.

Un exemplo actual que ilustra isto é a pandemia que estamos a vivir do Covid-19. Supoñamos que queremos estudar o tempo de duración do virus nos distintos habitantes dunha poboación. Tomamos como variable X o tempo que perdura a enfermidade nunha persoa. O estudo soamente terá información daquelas persoas que acudiron ao médico, quedando fóra do estudo aquelas que son asintomáticas ou se curaron na casa sen precisar de atención médica. Representamos por Y o tempo que transcurre dende que unha persoa presenta síntomas ata que acude ao médico. Como soamente temos constancia de aquelas que son observadas por un doutor, é equivalente a tomar aqueles datos que cumplen a restricción $Y \leq X$. Podemos ver ista situación gráficamente:



No primeiro capítulo deste traballo, o que faremos será unha estimación da distribución baixo o truncamento de datos. Comezaremos relacionando a distribución orixinal coa distribución que se obtén despois de truncar os datos, é dicir, de quedarnos só con aqueles que cumpran a condición $Y \leq X$, e continuaremos falando de estimadores e de diversas propiedades que presentan. Despois analizaremos tamén estimadores para o tamaño da mostra orixinal, N, así como análises de casos particulares que presentará o conxunto a risco dos datos observables. No segundo capítulo introduciremos varias simulacións feitas en R sobre diversos estimadores, comparándoos e analizando aqueles que sexan óptimos para os diferentes tipos de distribucións. Ao remate do traballo atópase un apéndice no cal se introducen unha serie de conceptos relacionados co truncamento, ao igual que as diferentes liñas de código de R usadas ao longo do traballo.

Capítulo 1

Estimación da distribución baixo truncamento

Este capítulo ímolo dedicar á búsqueda de estimadores da distribución orixinal da nosa poboación a partir dos n datos truncados. Para iso imos relacionar as distribucións antes e despois do truncamento, podendo estimar aquelas distribucións unha vez considerado o truncamento dos datos observados.

1.1. Relación entre a distribución orixinal e a truncada

O problema consistirá en dar unha estimación de F e G, que son as funcións de distribución marxinais de X e Y, respectivamente, así como da función de distribución conxunta sen truncamento de (X, Y), que denotaremos por $H = P(X \leq x, Y \leq y)$.

Definiremos como H_* á función de distribución conxunta de (X, Y) con truncamento, é dicir, á distribución de probabilidade de (X, Y) condicionada a que $Y \leq X$. H_* pode expresarse en función de F e G da seguinte maneira:

$$H_*(x,y) = P(X \leq x, Y \leq y / Y \leq X) = \frac{P(X \leq x, Y \leq y, Y \leq X)}{P(Y \leq X)} =$$

$$= \alpha^{-1} \int \int_{z \leq x, r \leq y, r \leq z} dF(z)dG(r) = \alpha^{-1} \int_{z \leq x} \int_{r \leq y, r \leq z} dG(r)dF(z) =$$

$$= \alpha^{-1} \int_0^x G(y \wedge z)\mathrm{d}F(z) \tag{1.1}$$

onde $y \wedge z$ denota o mínimo de y e z sendo $y \geq 0$ e $z < \infty$, e $\alpha = P(Y \leq X)$.

Por outro lado, α admite a expresión:

$$\alpha = P(Y \leq X) = \int_{y \leq x} dF(x)dG(y)$$

Podemos escribir a integral anterior de dúas maneiras, ben expresandoa como

$$\int_0^\infty G(z)dF(z)$$

ou como

$$\int_0^\infty (1 - F(z^-))dG(z)$$

sendo $F(z^-) = P(X < z)$.

Ademais, verifícanse as seguintes igualdades para as distribucións marxinais de X e Y con truncamento, respectivamente:

$$F_*(x) = H_*(x, \infty) \qquad 0 \leq x < \infty \tag{1.2}$$

$$G_*(y) = H_*(\infty, y) \qquad 0 \leq y < \infty \tag{1.3}$$

É difícil atopar estimadores consistentes para F e G de xeito que se satisfagan as condicións descritas na introducción, a menos que F_* e G_* determinen F e G. Vexamos que ocorre entón.

Sexa K calquera función de distribución en $[0, \infty]$, logo imos definir:

$$a_k = inf\{z > 0 : K(z) > 0\} \geq 0$$

$$b_k = sup\{z > 0 : K(z) < 1\} \leq \infty$$

onde (a_k, b_k) será o interior do soporte convexo de K, entendendo soporte convexo como o menor conxunto convexo que ten probabilidade un.

Temos que a_F, b_F son os extremos do soporte da nosa función F, mentras que a_G, b_G serano de G. Podémonos atopar situacións como as seguintes:

Cando $b_F < a_G$, non é observable ningún dato, como reflexa a gráfica da esquerda na Figura 1.1, onde a probabilidade da condición $Y \leq X$ é cero, i.e., $\alpha = P(Y \leq X) = 0$. Logo este tipo de casos non nos interesan.

Na gráfica da dereita da Figura 1.1, reflíctese unha situación na cal temos unha parte sombreada que non se observa, mentras que a outra parte dase construído xa que $\alpha = P(Y \leq X) > 0$. Traballaremos nas situacións deste tipo.

Sexa $\alpha > 0$ e F_* e G_* relacionadas con F e G por (1.1), (1.2) e (1.3), logo $a_{F_*} = max\{a_F, a_G\}$, $b_{F_*} = b_F$, $b_{G_*} = min\{b_F, b_G\}$ e $a_{G_*} = a_G$. Ademais, denotaremos os seguintes conxuntos:

$$\mathscr{K} = \{(F, G) : F(0) = 0 = G(0), \alpha(F, G) > 0\}$$

$$\mathscr{K}_0 = \{(F, G) \in \mathscr{K} : a_G \leq a_F, b_G \leq b_F\}$$

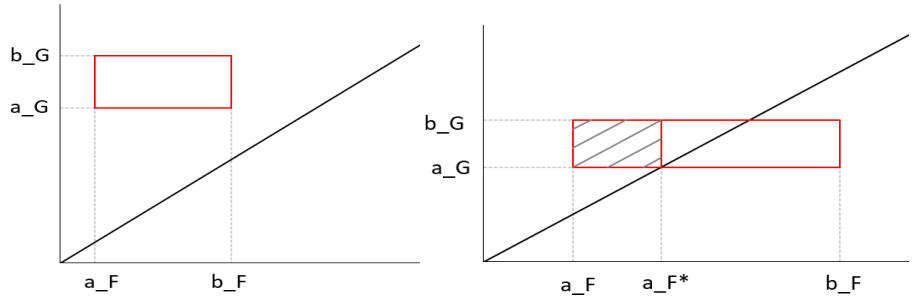


Figura 1.1: Representación de distintos tipos de conxuntos K.

$\mathscr{K}$ é un conxunto de distribucións de variables non negativas independentes onde existen algúns datos observables, mentras que $\mathscr{K}_0$ son distribucións onde todo o soporte é observable. Os pares pertencentes ao conxunto $\mathscr{K}_0$ denotarémolos por (F_0, G_0).

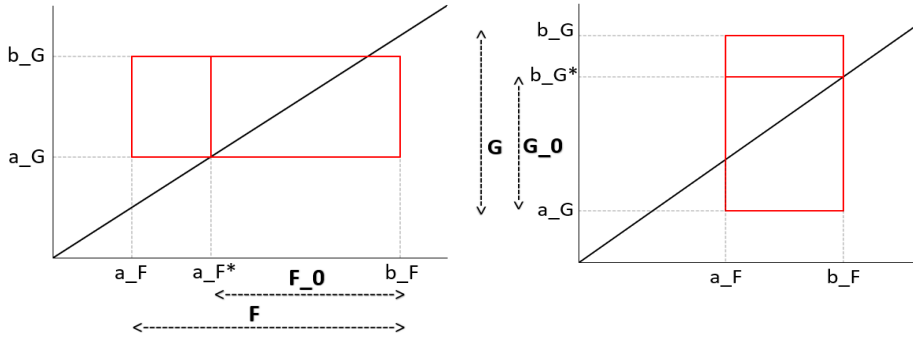


Figura 1.2: Representación de F e F_0 na gráfica da esquerda e de G e G_0 na da dereita.

F_0 determina a parte de F que podemos estimar, como se pode ver na Figura 1.2, xa que contén os datos observables, ao igual que ocorre con G_0.

Sexa T unha función a cal convirte a distribución orixinal na truncada, $T(F,G) = H_*$, $(F,G) \in \mathscr{K}$, polo tanto temos que $T(F_0,G_0) = T(F,G) = H_*$, sendo H_* a función de distribución conxunta da parte que podemos observar, como se ve sombreada na Figura 1.3, a cal estimaremos usando funcións empíricas.

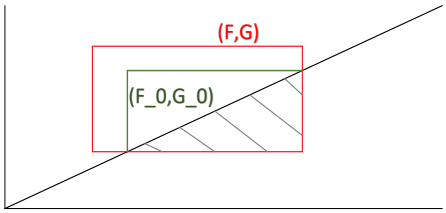


Figura 1.3: A zona sombreada representa o soporte da distribución truncada, H_*.

No seguinte teorema, expresaremos a función de risco acumulado de F, Λ_F, e a función de risco acumulado en tempo invertido de G, Λ_G^i en función de F_* e G_*. As funcións de risco e de risco en tempo invertido son funcións que caracterizan a distribución e que están explicadas no apéndice deste traballo.

Teorema 1.1. *Sexa $H_* \in T(\mathscr{K})$. Existe un único par $(F,G) \in \mathscr{K}_0$ tal que $T(F,G) = H_*$. Logo este par $(F,G) \in \mathscr{K}_0$ ven determinado por:*

$$\Lambda_F(x) = \int_0^x \frac{dF_*(z)}{C(z)} \qquad 0 \leq x < \infty \tag{1.4}$$

$$\Lambda_G^i(y) = \int_y^\infty \frac{dG(z)}{G(z)} = \int_y^\infty \frac{dG_*(z)}{C(z)} \qquad 0 \leq y < \infty \tag{1.5}$$

onde $C(z) = P(Y \leq z \leq X / Y \leq X) = G_*(z) - F_*(z^-)$ será denominado como o conxunto a risco.

A idea principal na que se basa a demostración do Teorema 1.1 é en expresar o conxunto a risco en función de F e G, substituíndo despois isto nas integrais (1.4) e (1.5) para obter así as expresións de $\Lambda_F(x)$ e $\Lambda_G^i(y)$. Escribimos $C(z)$ como:

$$\alpha C(z) = P(Y \leq X, Y \leq z) - P(Y \leq X, X < z) = P(Y \leq X, Y \leq z \leq X) =$$

$$= P(Y \leq z) - P(X < z, Y \leq z) = G(z) - G(z)F(z^-) = G(z)[1 - F(z^-)]$$

Polo tanto, para $0 \leq z < \infty$,

$$C(z) = \alpha^{-1} G(z)[1 - F(z^-)]$$

O que nos permite este teorema é estimar a función de risco acumulativo de X e a función de risco acumulativo en tempo invertido de Y, xa que ao lado dereito das igualdades atopámonos con funcións que dependen de F_* e G_*, as cales son estimables a través dos datos da mostras.

Corolario 1.2. *Sexa $(F, G) \in \mathscr{K}$ e sexan F_0 e G_0 as distribucións condicionais de X e Y dadas por $x \geq a_G$ e $y \leq b_F$. Logo, (F_0, G_0) é o único par en $\mathscr{K}_0$ tal que $T(F_0, G_0) = T(F, G)$.*

1.2. Estimadores das funcións de distribución

Como dixemos na introducción, imos tomar só aqueles pares (X_i, Y_i), $i \in \{1, ..., N\}$ que cumplan a condición $Y_i \leq X_i$. Supoñamos que sexan $(x_1, y_1), ..., (x_n, y_n)$, unha mostra de tamaño n. Empregando o Teorema 1.1 imos estimar F e G a partir de estimadores de F_* e G_*.

Describimos como F_n^* e G_n^* as funcións de distribución empíricas de $x_1, ..., x_n$ e $y_1, ..., y_n$, as cales estiman as funcións de distribución con truncamento, F_* e G_*, respectivamente. Veñen dadas por:

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq x)$$

$$G_n^*(y) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}(y_j \leq y)$$

onde $\mathbb{I}$ denota a función indicadora.

Polo Teorema 1.1, temos que a estimación da función de risco acumulado será:

$$\hat{\Lambda}_n(x) = \int_0^x \frac{dF_n^*(z)}{C_n(z)} = \sum_{x_i \leq x} \frac{1}{nC_n(x_i)} \qquad 0 \leq z < \infty \tag{1.6}$$

sendo

$$C_n(z) = G_n^*(z) - F_n^*(z^-)$$

o conxunto a risco estimado e $\hat{\Lambda}_n$, unha función escalonada con discontinuidades nos puntos $x_1, ..., x_n$.

Como consecuencia, asociado a esta razón de fallo discreta e usando o apéndice, obtemos as funcións estimadas de F e G:

$$\hat{F}_n(z) = 1 - \prod_{x_i \leq z} \left(1 - \frac{\sum_{k \leq n} \mathbb{I}(x_k = x_i)}{nC_n(x_i)}\right) \qquad 0 \leq z < \infty,\ 1 \leq i \leq n \tag{1.7}$$

$$\hat{G}_n(z) = \prod_{y_j > z} \left(1 - \frac{\sum_{k \leq n} \mathbb{I}(y_k = y_j)}{nC_n(y_j)}\right) \qquad 0 \leq z < \infty,\ 1 \leq j \leq n \tag{1.8}$$

Exemplo 1.3. Supoñamos o seguinte conxunto de datos:
(1,5),(1.5,1),(3,2.25),(2,4),(4,0.5),(4.5,3.5),(5,2.5),(2,1)
Na Figura 1.4 están representados para ver de forma máis clara como se distribuen.

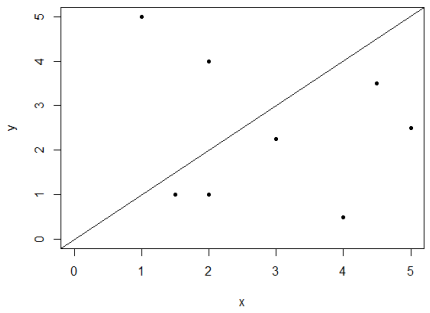


Figura 1.4: Representación dos datos da mostra.

Como dixemos anteriormente, quedarémonos con aqueles pares (x, y) que cumpren a condición $y \leq x$, que serán os datos observados a partir dos cales teremos que construír o resto da poboación. Na Figura 1.4 podemos ver que serían os datos debaixo da diagonal, sendo n=6 a mostra coa cal traballaremos. Imos explicar así a estimación da distribución F.

Podemos calcular entón, facendo uso do Teorema 1.1, o conxunto a risco de cada punto, $C_n(X_i)$, sendo $X_{(i)}$ cada un dos valores ordenados de X_i. O procedemento consiste en que a partir do individuo que cumple $Y \leq X$, trázase unha vertical ata a recta $X = Y$ e despois unha horizontal que sexa perpendicular a esa vertical polo punto de corte con $X = Y$. Desta maneira vanse debuxando recintos rectangulares como amosa a Figura 1.5. Os individuos os cales caen dentro deses recintos son os que corresponden cos distintos valores do conxunto a risco no instante X_i.

A partir desto, ilustraremos nunha táboa a razón de fallo estimada e as supervivencias de cada $X_{(i)}$, que virán dadas pola función de risco instantáneo, posto que se trata dunha

distribución discreta, xa que está baseada nun estimador empírico.

MOSTRA ORDEADA $X_{(i)}$	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$
RAZÓN DE FALLO ESTIMADA λ_i	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{2}$	1
SUPERVIVENCIA $\prod_{i\leq j}(1-\lambda_i)$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{4}{27}$	$\frac{2}{27}$	0

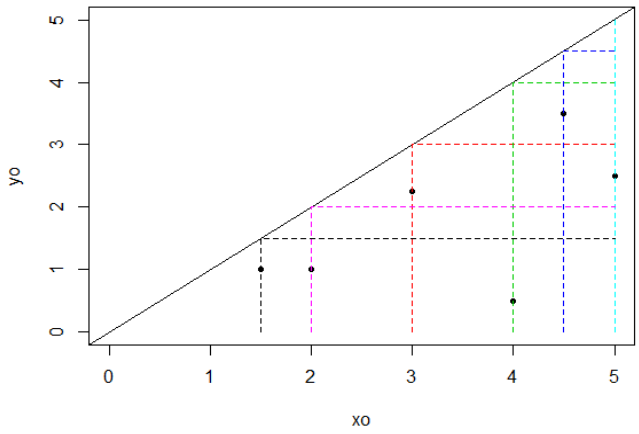


Figura 1.5: Conxuntos a risco do Exemplo 1.3.

Se a mostra non estivese truncada, teriamos que a probabilidade de fallo sería a mesma para todos os datos. Nese caso outorgaría unha probabilidade de $\frac{1}{6}$ a todos os individuos, debido que a mostra observable é $n = 6$. Podemos ver así, observando a táboa, que baixo truncamento a probabilidade de risco é maior e a probabilidade de supervivencia é menor que se a mostra non estivese truncada.

1.3. Conxuntos a risco unitarios

A continuación, falaremos dun caso particular que poden presentar os conxuntos a risco, chegando a traer problemas á hora de buscar un estimador para a distribución orixinal.

Sexan os estimadores $\hat{F}_n$ e $\hat{G}_n$. Escribimos os valores da mostra ordeada de X: $x_{(1)} < x_{(2)} < ... < x_{(n)}$.

Se algún conxunto a risco está formado por un único dato, i.e., $nC_n[x_{(k)}] = 1, 1 \leq k < n$, entón por (1.6) verificarase que $\hat{F}_n[x_{(k)}] = 1$, e queda probabilidade cero para todos os valores $x_{(i)}$ con $i > k$.

Incluso pode ocorrer que o conxunto a risco do primeiro dato da mostra ordeada sexa unitario, como mostra a Figura (1.6). Este caso é moi extremo e pode darse con máis probabilidade canto máis preto da orixe se atope ese punto. Polo tanto, teríamos que $\hat{F}_n[x_{(1)}] = 1$.

Isto fai que toda a probabilidade de supervivencia caia no primeiro dato, quedando os demáis datos sen probabilidade de sobrevivir. Isto pode chegar a parecer algo contradictorio, posto que o resto de individuos foron observados e sen embargo, non mostran supervivencia.

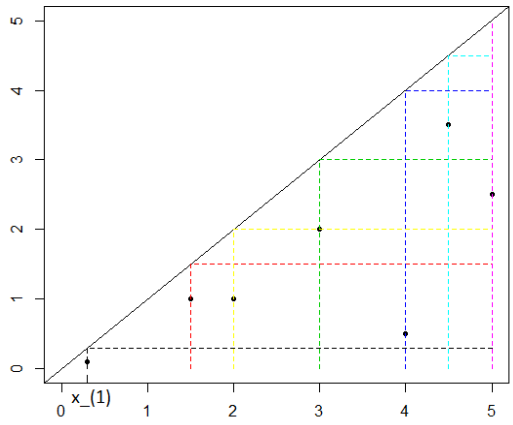


Figura 1.6: Conxunto a risco unitario do dato $x_{(1)}$.

Denotaremos por:

$$\Omega_o^{(n)} = \bigcup_{i=1}^{n-1} \{nC_n(x_{(i)}) = 1\}$$

á cantidade de conxuntos a risco unitarios que presenta cada mostra, sen ter en conta a última das observacións, posto que esta sempre presenta un conxunto a risco unitario. O artigo [3] fala destes ocos que presentan os conxuntos a risco, citando unha serie de resultados entre os cales podemos destacar o seguinte.

Lema 1.4. *Sexa F a función de distribución continua da variable aleatoria X. A probabilidade de atopar conxuntos a risco unitarios tende a cero cando aumenta o tamaño mostral, é dicir:*

$$\lim_{n\to\infty} P(\Omega_o^{(n)}) = 0.$$

Estas propiedades teóricas podemos velas nas simulacións do capítulo 2, na sección 2.2, onde se ilustrará con que frecuencia aparece algún conxunto a risco unitario.

Este tipo de problemas que presenta o feito de que $nC_n[x_{(k)}] = 1$, afectan á hora da construcción do estimador $\hat{F}_n$. Podemos resolvelos de distintas formas. Unha idea é incrementar lixeiramente os conxuntos a risco. Para iso tomaremos unha función k_n, a cal cumplirá que $k_n(x) > k_n[x_{(n)}] = \dfrac{1}{n} > 0$, para todo $x < x_{(n)}$. Así, remplazaremos o conxunto a risco por:

$$C'_n(z) = max\{C_n(z), k_n(z)\}, \quad 0 \leq z \leq x_{(n)}$$

na expresión (1.6) e (1.7), obtendo así $F'_n(z)$ e $G'_n(z)$, respectivamente, que non terán como soporte ningún subconxunto da mostra. Ademais, temos que $\dfrac{1}{nK_n[x_{(i)}]}$ será a máxima probabilidade estimada de $1 - F'_n[x_{(i)}]$ que se lle asignará a $x_{(i)}$. Esta función k_n o que fai é garantizar un número mínimo de datos en cada conxunto a risco.

1.4. Estimación do tamaño da poboación

Nesta sección imos falar de como se pode estimar N, que é o número de observacións da mostra orixinal do noso problema.

A partir dunha mostra observada de tamaño n e tendo un estimador de α, unha forma de estimar o tamaño da poboación orixinal será:

$$\hat{\alpha}_n \cdot \hat{N}_n = n \Rightarrow \hat{N}_n = \frac{n}{\hat{\alpha}_n}$$

Recordemos que:

$$\alpha = P(Y \leq X) = \int_{y \leq x} dF(x)dG(y) = \int_0^\infty G(z)dF(z)$$

Sabemos que $\hat{F}_n$ e $\hat{G}_n$ son estimadores das nosas funcións F e G, respectivamente, logo temos que o estimador de α será:

$$\hat{\alpha_n} = \int_0^\infty \hat{G}_n d\hat{F}_n \tag{1.9}$$

o cal será positivo se todos os conxuntos a risco non son unitarios, como veremos máis adiante.

Vexamos algúns exemplos nos cales teñamos que estimar o tamaño da poboación.

Exemplo 1.5. Faremos un estudo do número de persoas que se foron infectando por causa dunha enfermidade nunha certa localidade ao longo dos 5 primeiros días do mes de xuño. Tomamos como variable X o tempo (en días) que transcorre dende que unha persoa se infecta ata o día final do estudo, é dicir, ata o 5 de xuño; e como variable Y, o tempo

(en días) transcurrido dende que se infecta unha persoa ata que lle aparecen os primeiros síntomas. Supoñemos que (X, Y) son independentes, o que equivale a supoñer que o tempo dende a infección ata os síntomas non depende da data de infección, cousa que parece asumible.

Podemos ver, a continuación, nas Figuras (1.7) e (1.8), unha representación gráfica das situacións ante as cales poden presentarse as persoas. Na Figura (1.7), os datos serán observables xa que as persoas infectanse e presentan síntomas antes de que finalice o estudo, é dicir, cumpren a condición $Y \leq X$. Sen embargo, na Figura (1.8), estes datos non serán observados xa que $Y > X$. Isto pode ser o caso dunha persoa que se infecta pero non padece síntomas ata despois do día 5 de xuño, polo tanto, estes datos non serán tidos en conta á hora de facer o estudo.

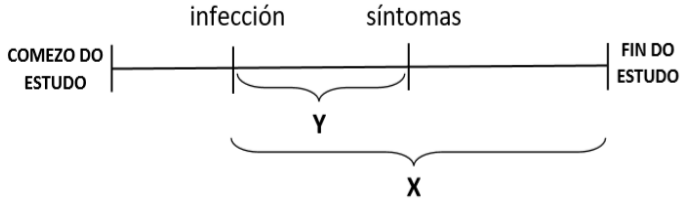


Figura 1.7: Persoas que serán observadas xa que $Y \leq X$.

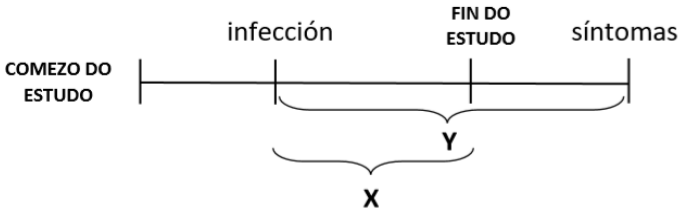


Figura 1.8: Persoas que non serán observadas xa que $Y > X$.

No Cadro 1.1 temos recollidos o número de persoas que se foron infectando durante estes 5 días (datos ficticios), tendo só constancia de aqueles individuos que se atopan nas celdas grises, xa que son os datos que cumpren a condición de truncamento $Y \leq X$, e polo tanto serán os datos observados. Neste caso temos $n = 160$, xa que sería a suma de todos os individuos das celdas observadas. A partir destes n individuos observados, estimaremos os datos truncados e aproximaremos así a poboación orixinal de infectados.

	X	Y				
		0	1	2	3	4
1 xuño	4	8	12	11	7	9
2 xuño	3	9	13	12	10	* 10,42
3 xuño	2	12	14	12	** 9,94	*** 11,35
4 xuño	1	10	9	9,78	7,53	8,6
5 xuño	0	12	14,77	13,78	10,61	12,12

Cadro 1.1: Persoas que se infectaron durante os 5 primeiros días do mes de xuño. Na parte superior da escaleira as que padeceron síntomas antes do día 5 e na parte inferior, as que padeceron síntomas despois do día 5 (datos ficticios).

Observamos que o número de persoas as cales contraeron a enfermidade o día 1 de xuño suman un total de 47, a pesar de presentar síntomas en distintos días. No Cadro 1.1 vemos que foron 8 as persoas que se infectaron o día 1 de xuño e que presentaron síntomas ese mesmo día, sendo, por exemplo, 12, as persoas as cales se infectaron tamén o 1 de xuño pero empezaron a ter síntomas un día despois. Na seguinte fila podemos observar que o día 2 de xuño infectáronse e presentaron síntomas o mesmo día 9 persoas.

Ademais, podemos ver que se unha persoa se infecta o día 2 de xuño e comeza a ter síntomas pasados 4 días, sería día 6 de xuño, e supoñemos que estamos a día 5, polo que non temos constancia desta persoa. É o que se corresponde coa celda co asterisco $(*)$ do Cadro 1.1. O que faremos a continuación será estimar a cantidade de individuos nesta celda cunha simple regra de tres (regra de proporcionalidade):

$$8 + 12 + 11 + 7 = 38 \quad \text{——} \quad 9$$

$$9 + 13 + 12 + 10 = 44 \quad \text{——} \quad x$$

Polo tanto temos que $x = \dfrac{44 \cdot 9}{38} = 10{,}42$ persoas foron infectadas 4 días despois do 2 de xuño, é dicir, o día 6.

Agora faremos unha estimación das persoas infectadas o día 3 de xuño que non tiveron síntomas ata 3 e 4 días despois, é dicir, ata o día 6 e 7 de xuño, respectivamente. Estas celdas no Cadro 1.1 están representadas por $(**)$ e $(***)$.

$$8 + 12 + 11 + 9 + 13 + 12 = 65 \quad \text{——} \quad 17 = 7 + 10$$

$$12 + 14 + 12 = 38 \quad \text{——} \quad x$$

Polo tanto neste caso temos a $x = \dfrac{38 \cdot 17}{65} = 9{,}94$ persoas que se infectaron o día 3 e padeceron síntomas o día 6. Coa axuda deste novo dato, imos ver aquelas que comezaron a presentar síntomas o día 7 de xuño, que se corresponden coa celda $(***)$.

$$8 + 12 + 11 + 7 + 9 + 13 + 12 + 10 = 82 \quad \text{———} \quad 19{,}42 = 9 + 10{,}42$$

$$12 + 14 + 12 + 9{,}94 = 47{,}94 \quad \text{———} \quad x$$

Temos así $x = \dfrac{47{,}94 \cdot 19{,}42}{82} = 11{,}35$ persoas.

Continuaríamos facendo isto ata conseguir completar a táboa. Podemos chegar así, mediante estes cálculos, a unha estimación daqueles datos que estaban truncados, e en consecuencia, a unha estimación do tamaño total da poboación, $\hat{N}_n = 268{,}9$. Ás veces pode ocorrer que as aproximacións destes valores sexan distantes aos datos orixinais da poboación debido ao azar do tempo entre a infección e os síntomas, xa que non ten que presentar un patrón idéntico en todas as datas de infeccións.

Podemos plantexarnos a continuación unha situación na cal teriamos que estimar o tamaño da poboación, N, e estaría presente o problema dos conxuntos a risco unitarios do cal falabamos na sección 1.3.

Imos tomar a mesma situación do Exemplo 1.4; as persoas infectadas debido a unha enfermidade nunha poboación, pero esta vez só faremos o estudo durante unha fin de semana.

Exemplo 1.6. Sexa $(X_1, Y_1), ..., (X_N, Y_N)$ unha poboación. Tomaremos a variable aleatoria X como o tempo (en días) dende que se infecta unha persoa ata que remata a fin de semana e a variable aleatoria Y como o tempo (en días) que pasa dende que unha persoa se contaxia ata que presenta síntomas.

O estudo só durará o sábado e o domingo, polo tanto soamente se contarán aquelas persoas que contraeron a enfermidade o sábado e a padeceron ese mesmo día ou ao día seguinte; máis aquelas que se infectaron e deron síntomas o domingo. Esto é equivalente a tomar aqueles datos que cumpran a condición $Y \leq X$.

Existirán persoas que se infectaron o domingo e cuxos síntomas non lles aparecerán ata o luns. O dato que representa ao número de persoas nesta situación estará truncado, xa que está fóra do estudo, debido a que cumpre a condición $Y > X$.

Daquela, hai tres celdas observadas, as cales vemos amosadas no Cadro 1.2.

	X	Y	
		0	1
Sábado	1	30	15
Domingo	0	20	?

Cadro 1.2: Persoas que se infectaron e padeceron a enfermidade o mesmo día ou ao día seguinte (datos ficticios).

O dato truncado podemos calculalo co mesmo procedemento que usamos no Exemplo 1.4, resultando 10 como estimación do número de persoas infectadas o domingo presentando síntomas o luns.

Supoñamos estas mesmas variables na situación do Cadro 1.3, é dicir, non existe ningunha persoa que se infectara o sábado e padecera síntomas ese mesmo día. Esto provoca a existencia dun conxunto a risco dexenerado, no sentido de que todos os individuos caen na celda do risco. Polo tanto, neste caso resulta complexo calcular o dato truncado.

	X	Y	
		0	1
Sábado	1	0	15
Domingo	0	20	?

Cadro 1.3: Persoas que se infectaron e padeceron a enfermidade o mesmo día ou ao día seguinte. Existencia dun conxunto a risco dexenerado.

Nesta situación temos que $\hat{\alpha}_n = 0$ e non é posible estimar N. Así, cando aparece un conxunto a risco unitario (dexenerado en táboas discretas) non se pode estimar N, pero ademais, non sendo unitario, cando é pequeno provoca situacións desorbitadas de N.

Este Exemplo 1.6 ilustrounos que hai situacións nas cales pode resultar complexo o feito de facer estimacións, xa que os estimadores tamén poden plantexar defectos debido ao azar. No capítulo 2 deste traballo faremos simulacións e veremos a probabilidade que existe de que no conxunto de datos tomado aparezan conxuntos a risco unitarios.

1.5. Propiedades teóricas dos estimadores: Consistencia e converxencia da distribución

A consistencia dun estimador está relacionada co comportamento que este experimenta cando o tamaño da mostra aumenta, facendo así que a un maior tamaño de mostra se proporcione máis información sobre a poboación orixinal.

Tomaremos F e G funcións de distribución continuas de X e Y, respectivamente, tales que $(F, G) \in \mathscr{K}$, e sexa $(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)$ a poboación dada, con tamaño N.

Consideremos os estimadores $\hat{F}_n$ e $\hat{G}_n$ calculados a partir das expresións (1.7) e (1.8). Estudaremos o comportamento que presentan os estimadores anteriores cando $N \to \infty$.

Sexan $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ valores independentes que cumpren a condición de truncamento $Y_i \leq X_i$. Estes valores seguen a función de distribución conxunta con truncamento, H_*.

Ao cardinal do conxunto $\{i \leq N : Y_i \leq X_i\}$, con $N > 0$, denótaselle por n. Logo n segue unha distribución binomial con $N > 0$ intentos e probabilidade de éxito $P(Y \leq X) = \alpha$, é dicir, $n \sim B(N, \alpha)$.

Denotamos entón por P_n á probabilidade condicionada sabendo que hai n datos observados. As converxencias de $\hat{F}_n$ e $\hat{G}_n$ en probabilidade P_n están determinadas cando $n \to \infty$, e en consecuencia, serán tamén as converxencias na probabilidade non condicionada cando $N \to \infty$.

A continuación, enunciaremos algúns resultados relacionados coa converxencia uniforme, os cales están probados en documentos como [1].

Teorema 1.7. *Sexan F e G funcións de distribución continuas para as cales $(F, G) \in \mathscr{K}$ e denotamos por F_0 e G_0 as funcións de distribución condicionadas de X e Y tales que $(F_0, G_0) \in \mathscr{K}_0$. Polo tanto, temos que:*

$$\sup_{x>0} |\hat{F}_n(x) - F_0(x)| \longrightarrow 0$$

$$\sup_{y>0} |\hat{G}_n(y) - G_0(y)| \longrightarrow 0$$

en probabilidade P_n cando $n \to \infty$.

Corolario 1.8. *Se F e G son continuas e temos o par $(F_0, G_0) \in \mathscr{K}_0$, logo cúmprese que:*

1. *$\hat{\alpha}_n \longrightarrow \alpha$ en probabilidade P_n cando $n \to \infty$.*
2. *$\frac{\hat{N}_n}{N} \longrightarrow 1$ en probabilidade cando $N \to \infty$.*

Ademáis, en [1, Sección 5 e 6] establécese tamén a converxencia en distribución de $\hat{F}_n$ a F.

Capítulo 2

Simulacións en R

Neste capítulo imos facer varios estudos de simulación. Na sección 2.1 analizaremos as propiedades do estimador da media, como poden ser o sesgo, a varianza e o erro cadrático medio.

Na sección 2.2 as simulacións centraranse nos conxuntos a risco, facendo unha análise dos problemas que estes poden supoñer.

Finalmente, na sección 2.3, imos estudar as propiedades do estimador do tamaño da poboación total a partir da mostra observada.

2.1. Estimación da media

Consideremos $(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)$. O noso problema consiste en estimar a media da variable X na poboación orixinal, $\mathbb{E}(X) = \mu$. Tomaremos aquelas observacións que cumpren a condición de truncamento, $Y_i \leq X_i$, e denotarémolas por $(x_1, y_1), ..., (x_n, y_n)$. Escribiremos dous posibles estimadores para a estimación de μ: a media aritmética simple das observacións e a media ponderada polos saltos do estimador $\hat{F}_n$.

$$\hat{\mu}_o = \frac{1}{n} \sum_{i=1}^{n} x_{(i)} \tag{2.1}$$

$$\hat{\mu}_T = \sum_{i=1}^{n} w_i x_{(i)} = \sum_{i=1}^{n} (\hat{F}_n(x_{(i)}) - \hat{F}_n(x_{(i-1)}))x_{(i)} \tag{2.2}$$

onde $x_{(i)}$, $i \in \{1, ..., n\}$, denota o elemento que ocupa a posición i-ésima na mostra ordeada.

Como se trata de variables aleatorias, podemos calcularlles a esperanza, $\mathbb{E}(\hat{\mu}_o)$ e $\mathbb{E}(\hat{\mu}_T)$, e en consecuencia, o sesgo, que é a diferenza entre a esperanza matemática do estimador e o valor numérico do parámetro que se está a estimar,

$$sesgo(\hat{\mu}_o) = \mathbb{E}(\hat{\mu}_o) - \mu \qquad sesgo(\hat{\mu}_T) = \mathbb{E}(\hat{\mu}_T) - \mu$$

Podemos ademais, calcular as varianzas dos respectivos estimadores, $Var(\hat{\mu}_o)$ e $Var(\hat{\mu}_T)$, e achar así os seus erros cadráticos medios, que se definen como o promedio dos erros cadráticos ao cadrado e coinciden coa suma que se obtén do sesgo ao cadrado e a varianza do estimador,

$$ECM(\hat{\mu}_o) = \mathbb{E}[(\hat{\mu}_o - \mu_o)^2] = sesgo(\hat{\mu}_o)^2 + Var(\hat{\mu}_o)$$

$$ECM(\hat{\mu}_T) = \mathbb{E}[(\hat{\mu}_T - \mu_T)^2] = sesgo(\hat{\mu}_T)^2 + Var(\hat{\mu}_T)$$

O estimador $\hat{\mu}_o$ é un estimador da media máis sinxelo, sen embargo este non ten en conta o truncamento dos datos. Realmente, $\hat{\mu}_o$ é un estimador da esperanza condicionada, $\mathbb{E}(X/Y \leq X)$, mentras que $\hat{\mu}_T$ é un estimador adecuado da esperanza, $\mathbb{E}(X)$, que ten en conta o truncamento.

O que faremos a continuación será estimar as propiedades de $\hat{\mu}_o$ e $\hat{\mu}_T$ mediante simulacións en R. Consideremos S o número total de mostras simuladas. Con cada mostra obtense un valor do estimador, resultando así S valores de $\hat{\mu}_o$ e outros S valores de $\hat{\mu}_T$.

Denotaremos entón por $\hat{\mu}_o^1, \hat{\mu}_o^2, ..., \hat{\mu}_o^S$ aos S valores de $\hat{\mu}_o$, e por $\hat{\mu}_T^1, \hat{\mu}_T^2, ..., \hat{\mu}_T^S$ aos S valores de $\hat{\mu}_T$. Con eles podemos aproximar os sesgos, varianzas e erros cadráticos medios.

Para $\hat{\mu}_o$ sería:

$$\widehat{\mathbb{E}}(\hat{\mu}_o) = \frac{1}{S}\sum_{s=1}^{S} \hat{\mu}_o^s \tag{2.3}$$

$$\widehat{sesgo}(\hat{\mu}_o) = \widehat{\mathbb{E}}(\hat{\mu}_o) - \mu \tag{2.4}$$

$$\widehat{Var}(\hat{\mu}_o) = \frac{1}{S-1}\sum_{s=1}^{S} (\hat{\mu}_o^s - \widehat{\mathbb{E}}(\hat{\mu}_o))^2 \tag{2.5}$$

$$\widehat{ECM}(\hat{\mu}_o) = \widehat{sesgo}(\hat{\mu}_o)^2 + \widehat{Var}(\hat{\mu}_o) \tag{2.6}$$

e para $\hat{\mu}_T$:

$$\widehat{\mathbb{E}}(\hat{\mu}_T) = \frac{1}{S}\sum_{s=1}^{S} \hat{\mu}_T^s \tag{2.7}$$

$$\widehat{sesgo}(\hat{\mu}_T) = \widehat{\mathbb{E}}(\hat{\mu}_T) - \mu \tag{2.8}$$

$$\widehat{Var}(\hat{\mu}_T) = \frac{1}{S-1}\sum_{s=1}^{S} (\hat{\mu}_T^s - \widehat{\mathbb{E}}(\hat{\mu}_T))^2 \tag{2.9}$$

$$\widehat{ECM}(\hat{\mu}_T) = \widehat{sesgo}(\hat{\mu}_T)^2 + \widehat{Var}(\hat{\mu}_T) \tag{2.10}$$

A continuación imos ilustrar todo isto cuns exemplos, onde tomaremos distintas distribucións e faremos nelas unha análise dos estimadores da media $\hat{\mu}_o$ e $\hat{\mu}_T$.

Exemplo 2.1. Sexa $(X_1, Y_1), ..., (X_N, Y_N)$ unha poboación xerada no programa R. Consideremos F e G distribucións expoñenciais con parámetros $\lambda_X = \frac{1}{3}$, $\lambda_Y = \frac{1}{3}$, respectivamente, sendo λ_X e λ_Y as razóns de fallo do modelo expoñencial. Denotaremos o tamaño da poboación total por N. Os valores de dita poboación que son observables, xa que cumpren a condición $Y \leq X$, serán $(x_1, y_1), ..., (x_n, y_n)$, tendo así unha mostra de n datos observados. Quedarémonos con esta mostra de tamaño n e iremos analizando os estimadores para a media (2.1) e (2.2) en distintas situacións.

Imos reproducir, por exemplo, $S = 10000$ simulacións, e imos fixar inicialmente o tamaño da poboación, N. A partir deste, en cada unha das distintas simulacións imos obter un número aleatorio de datos observables, n. Teremos entón mostras do tipo:

$$(x_1^1, y_1^1), (x_2^1, y_2^1), ..., (x_{n_1}^1, y_{n_1}^1)$$

$$\vdots$$

$$(x_1^S, y_1^S), (x_2^S, y_2^S), ..., (x_{n_S}^S, y_{n_S}^S)$$

Coa axuda de R, obtivemos os seguintes valores dos estimadores (2.1) e (2.2).
Para $N = 60$:

$$\hat{\mu}_o : 5.506,\ 4.549,...,\ 3.966$$
$$\hat{\mu}_T : 3.543,\ 3.480,...,\ 2.487$$

Para $N = 100$:

$$\hat{\mu}_o : 5.360,\ 4.854,...,\ 3.816$$
$$\hat{\mu}_T : 3.778,\ 3.003,...,\ 2.759$$

Para $N = 300$:

$$\hat{\mu}_o : 4.720,\ 4.272,...,\ 4.652$$
$$\hat{\mu}_T : 3.591,\ 2.678,...,\ 3.487$$

Sabemos que a media teórica dunha distribución expoñencial é $\frac{1}{\lambda}$. Polo tanto, no noso caso temos que:

$$\mathbb{E}[X] = \frac{1}{\lambda_x} = \frac{1}{\frac{1}{3}} = 3.$$

Con estes datos, e usando as expresións (2.4), (2.5), (2.6), (2.8), (2.9) e (2.10), podemos obter os resultados recollidos por columnas, respectivamente, no Cadro 2.1.

N	$\widehat{sesgo}(\hat{\mu}_o)$	$\widehat{var}(\hat{\mu}_o)$	$\widehat{ecm}(\hat{\mu}_o)$	$\widehat{sesgo}(\hat{\mu}_T)$	$\widehat{var}(\hat{\mu}_T)$	$\widehat{ecm}(\hat{\mu}_T)$
60	1.5	0.374	2.625	0.054	0.734	0.737
100	1.5	0.225	2.475	0.034	0.477	0.478
300	1.5	0.075	2.323	0.010	0.192	0.192

Cadro 2.1: Estimacións de sesgo, varianza e erro cadrático medio dos estimadores da media fixado N.

O que se fixo foi ir estimando as distintas propiedades das variables aleatorias, aumentando o tamaño da poboación, N, para ver como varían os resultados.

Podemos calcular o sesgo exacto de $\hat{\mu}_o$ mediante unha serie de cálculos como indicaremos a continuacón, xa que se corresponden cos datos observables da mostra dos cales temos información. No Cadro 2.1 os valores que reflexa non son exactos, son resultados dunha simulación, pero vemos que coinciden polo feito de efectuar 10000 mostras simuladas.

Calcularemos entón $\mathbb{E}(\hat{\mu}_o)$, para obter así o valor do sesgo de $\hat{\mu}_o$.

$$\mathbb{E}(\hat{\mu}_o) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_{(i)}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i/Y_i \leq X_i) = \frac{1}{n}\cdot n\cdot\mathbb{E}(X_i/Y_i \leq X_i) = \mathbb{E}(X_i/Y_i \leq X_i)$$

Por outra banda, facendo uso da expresión (1.2), chegamos a que:

$$\mathbb{E}(X/Y \leq X) = \int_0^\infty x dF_*(x) = \int_0^\infty x\alpha^{-1}G(x)dF(x) = \alpha^{-1}\int_0^\infty xG(x)dF(x)$$

Temos que a probabilidade de observación é $\frac{1}{2}$, é dicir, $\alpha = P(Y \leq X) = \frac{1}{2}$, porque tanto X como Y presentan a mesma distribución.

Por outra banda sabemos que $\lambda_X = \frac{1}{3} = \lambda_Y$, obtendo así as seguintes funcións:

$$g(x) = \lambda_Y e^{-\lambda_Y x} = \frac{1}{3}e^{-\frac{x}{3}} \Rightarrow G(x) = 1 - e^{-\frac{x}{3}}$$

$$dF(x) = f(x)dx = \lambda_X e^{-\lambda_X x}dx = \frac{1}{3}e^{-\frac{x}{3}}dx$$

Substituíndo todo isto na expresión anterior e facendo unha serie de cálculos:

$$\mathbb{E}(X/Y \leq X) = 2\int_0^\infty x(1-e^{-\frac{x}{3}})\frac{1}{3}e^{-\frac{x}{3}}dx = 2[\int_0^\infty \frac{1}{3}xe^{-\frac{x}{3}}dx - \int_0^\infty \frac{1}{3}xe^{-\frac{2x}{3}}dx] =$$

$$= 2\int_0^\infty \frac{1}{3}xe^{-\frac{x}{3}}dx - 2\int_0^\infty \frac{1}{3}xe^{-\frac{2x}{3}}dx = ... = 2\cdot 3 - 2\cdot\frac{3}{4} = \frac{9}{2} = 4{,}5$$

Temos, polo tanto:

$$Sesgo(\hat{\mu}_o) = \mathbb{E}(\hat{\mu}_o) - \mu = 4{,}5 - 3 = 1{,}5$$

Vemos que o estimador $\hat{\mu}_o$ non proporciona boas estimacións da media de X debido a que o sesgo non converxe a cero, senón que se mantén alto, como podería ser fácil de intuír, xa que este estimador só está a traballar cos datos observables e non coa mostra na súa totalidade. De feito, como estimador de $\mathbb{E}(X/Y \leq X)$ sería un estimador insesgado.

Pola contra, no caso do estimador $\hat{\mu}_T$, o sesgo é moi baixo, vendo que a medida que aumenta o tamaño da poboación, tende cada vez máis a cero, dando unha mellor aproximación da media.

Por outro lado, temos a varianza dos estimadores $\hat{\mu}_o$ e $\hat{\mu}_T$, nas columnas 3 e 6 do Cadro 2.1, respectivamente. A varianza é unha medida de dispersión non negativa, a cal é cero se todos os datos son iguais e vai aumentando a medida que os datos están máis dispersos. Podemos observar que a medida que aumenta o tamaño da poboación, o número de datos será maior, e polo tanto os resultados do estimador serán mellores, facendo que a varianza se vaia achegando a cero, como ocorre cos resultados que devolven os estimadores $\hat{\mu}_T$ e $\hat{\mu}_o$.

Sen embargo, vemos que a varianza de $\hat{\mu}_T$ é maior que a de $\hat{\mu}_o$. Isto é debido a que $\hat{\mu}_T$ procede dun mecanismo complicado de corrección do truncamento, mentras que $\hat{\mu}_o$ é unha media aritmética simple.

Finalmente, podemos observar no Cadro 2.1 os erros cadráticos medios dos estimadores. O erro cadrático medio o que nos dará será a media das diferenzas ao cadrado entre o estimador e o valor real. Canto menor sexa este erro, mellor será o estimador.

Vemos que a medida que aumenta o número de datos da poboación, existirá unha menor dispersión destes datos e en consecuencia, un mellor axuste do estimador que estamos a analizar. Podemos dicir que o estimador μ_T é consistente, posto que o seu erro cadrático medio tende a cero cando N tende a ∞ e polo tanto dará unha boa aproximación da media.

O estudo que fixemos ata agora foi fixando o tamaño da poboación. O que faremos a continuación será fixar o número de datos observables, n, e faremos unha análise similar a anterior. Teremos que extraer unha cantidade aleatoria N de pares (X, Y), pois tomaremos os que sexan necesarios ata chegar a n pares cumprindo a condición $Y \leq X$. Reproduciremos tamén $S = 10000$ mostras simuladas. Teremos entón para $n = 30$:

$$(x_1^1, y_1^1), (x_2^1, y_2^1), ..., (x_{30}^1, y_{30}^1)$$
$$\vdots$$
$$(x_1^S, y_1^S), (x_2^S, y_2^S), ..., (x_{30}^S, y_{30}^S)$$

sendo os estimadores $\hat{\mu_o}$ e $\hat{\mu_T}$:

$$\hat{\mu}_o : 5.582,\ 4.708,...,\ 5.137$$
$$\hat{\mu}_T : 3.637,\ 3.236,...,\ 4.110$$

Do mesmo xeito ocorrerá, por exemplo, co tamaño de mostra observable $n = 50$:

$$\hat{\mu}_o : 5.287,\ 5.081,...,\ 4.783$$
$$\hat{\mu}_T : 3.718,\ 3.098,...,\ 2.000$$

e co tamaño $n = 150$:

$$\hat{\mu}_o : 4.921,\ 4.533,...,\ 3.961$$
$$\hat{\mu}_T : 3.354,\ 2.518,...,\ 2.729$$

No Cadro 2.2 están escritos os valores resultantes das aproximacións do sesgo, varianza e erro cadrático medio dos estimadores μ_o e μ_T para os diferentes n.

n	$\widehat{sesgo}(\hat{\mu}_o)$	$\widehat{var}(\hat{\mu}_o)$	$\widehat{ecm}(\hat{\mu}_o)$	$\widehat{sesgo}(\hat{\mu}_T)$	$\widehat{var}(\hat{\mu}_T)$	$\widehat{ecm}(\hat{\mu}_T)$
30	1.5	0.360	2.6	0.039	0.725	0.727
50	1.5	0.222	2.461	0.025	0.475	0.476
150	1.5	0.075	2.332	0.009	0.189	0.189

Cadro 2.2: Estimacións de sesgo, varianza e erro cadrático medio dos estimadores da media fixado n.

Imos comparar entón o Cadro 2.1 co Cadro 2.2, facendo unha análise das diferenzas que ocorren cando fixamos N ou n. Podemos observar que os valores de n fixados no Cadro 2.2 son xustamente a metade dos valores de N fixados no Cadro 2.1, pois $\alpha = P(Y \leq X) = 0{,}5$, tomados así para ver como inflúe o fenómeno do azar de n cando fixamos N.

Podemos pensar que os resultados para $N = 60$ do Cadro 2.1 deberían ser os mesmos ou similares aos resultados que presenta o Cadro 2.2 para o valor $n = 30$, se en todas as simulacións se presentaran 30 datos observados, pero aquí entra en xogo o factor da aleatoriedade de n. Este é o feito polo cal os resultados que se obteñen no Cadro 2.1 son distintos aos que se obteñen no Cadro 2.2.

Podemos concluír tamén, á vista de ambas táboas e debido a este factor da aleatoriedade, que o estimador $\hat{\mu}_T$ presenta unha milloría no sesgo e na varianza cando fixamos n respecto da situación en que fixamos N.

Exemplo 2.2. A continuación faremos un estudo similar ao do Exemplo 2.1, pero seguiremos unha distribución uniforme no intervalo [0,1]. Consideramos F e G as funcións marxinais e denotaremos por N o tamaño total da poboación. Os valores observables, é dicir, os que cumpren a condición $Y \leq X$, terán tamaño n, $(x_1, y_1), ...(x_n, y_n)$. Quedarémonos con esta mostra e iremos analizando os estimadores (2.1) e (2.2).

Recordemos que a media teórica nunha distribución uniforme vén dada por:

$$\mathbb{E}[X] = \frac{a+b}{2}$$

sendo no noso caso $\mathbb{E}[X] = \frac{1}{2}$.

Imos xerar entón $S = 10000$ simulacións e imos fixar o número de datos observables, n.

Para $n = 40$:

$$\hat{\mu}_o : 0.607,\ 0.663,...,\ 0.679$$
$$\hat{\mu}_T : 0.419,\ 0.473,...,\ 0.518$$

Para $n = 120$:

$$\hat{\mu}_o : 0.661,\ 0.654,...,\ 0.701$$
$$\hat{\mu}_T : 0.514,\ 0.483,...,\ 0.603$$

Para $n = 200$:

$$\hat{\mu}_o : 0.667,\ 0.667,...,\ 0.648$$
$$\hat{\mu}_T : 0.504,\ 0.537,...,\ 0.521$$

Construiremos entón a seguinte táboa:

n	$\widehat{sesgo}(\hat{\mu}_o)$	$\widehat{var}(\hat{\mu}_o)$	$\widehat{ecm}(\hat{\mu}_o)$	$\widehat{sesgo}(\hat{\mu}_T)$	$\widehat{var}(\hat{\mu}_T)$	$\widehat{ecm}(\hat{\mu}_T)$
40	0.167	0.001	0.029	0.007	0.011	0.0.011
120	0.167	0.0005	0.028	0.002	0.005	0.005
200	0.167	0.0003	0.028	0.001	0.003	0.003

Cadro 2.3: Estimacións de sesgo, varianza e erro cadrático medio de estimadores da media fixado n.

Vemos, neste caso, que o sesgo do estimador $\hat{\mu}_o$ é 0.167. O valor 0.167 pódese achar facendo uns cálculos similares aos do Exemplo 2.1.

Na columna 5 podemos observar que o valor do sesgo do estimador $\hat{\mu}_T$ disminúe ao aumentar o tamaño de n, tendendo a cero, como ocurría no Exemplo 2.1, sendo así este estimador asintóticamente insesgado.

Respecto á dispersión que existe entre os datos, vemos que ambos estimadores presentan unha varianza que tende a cero a medida que os datos observados aumentan (columnas 3 e 6 do Cadro 2.3).

Finalmente, podemos falar tamén do erro cadrático medio de $\hat{\mu}_o$ e $\hat{\mu}_T$, sabendo que canto máis pequeno sexa este, o estimador dará unha mellor aproximación da media. Vemos que a columna 4 do Cadro 2.3 presenta un erro cadrático medio maior que o que reflexa a columna 7. De feito, o erro cadrático medio de $\hat{\mu}_o$ con n grande redúcese ao seu sesgo ao cadrado que se mantén constante en $(0{,}167)^2 = 0{,}028$.

2.2. Estimación de conxuntos a risco unitarios

A continuación trataremos co problema que se nos presentaba na sección 1.3 deste traballo, onde se falaba sobre a presenza de conxuntos a risco unitarios. O dato que pertenza a este conxunto a risco unitario fai que toda a probabilidade de sobrevivir recaiga nel nese instante x_i, deixando así ao resto de datos da súa dereita con probabilidade cero.

Exemplo 2.3. Sexa $(X_1, Y_1), ..., (X_N, Y_N)$ unha poboación xerada no programa *R*. Consideremos F e G distribucións expoñenciais con parámetros $\lambda_X = \frac{1}{3}$, $\lambda_Y = \frac{1}{3}$, respectivamente. Tomamos unha mostra ordeada de tamaño n, cumprindo a condición de truncamento $Y \leq X$: $(x_{(1)}, y_{[1]}), ..., (x_{(n)}, y_{[n]})$, sendo $y_{[i]}$ os concomitantes que acompañan ao estatístico ordeado de x na posición i con $i \in \{1, ..., n\}$.

O truncamento o que produce é unha perda de datos da poboación total, o cal afecta á hora de calcular os conxuntos a risco de cada observación. Se non existira truncamento, todos os datos serían observados, presentando así conxunto a risco sen perdas de información.

Os conxuntos a risco "pequenos" poden xerar problemas, sobre todo se se atopan ao principio da mostra. Imos poñer un exemplo do que ocorre cunha primeira mostra simulada de tamaño $n = 45$. Vemos que no conxunto a risco do primeiro dato da primeira mostra, aparece un 8. Outorgaráselle así unha probabilidade de fallo a $x_{(1)}$ de $\frac{1}{8}$ e unha probabilidade de supervivencia de $\frac{7}{8}$, sendo esta supervivencia moi pequena para ser un dato dos primeiros.

```
> cn
[1]  8 10 20 19 20 22 21 21 22 23 24 25 24 24 25 24 23 22 23 22 21 21 20 19
[25] 19 19 18 17 16 15 14 13 12 11 11 10  9  8  7  6  5  4  3  2  1
```

Supoñamos agora que xeramos $S = 1000$ mostras simuladas para un tamaño fixado de n observacións. Podemos estudar a distribución dos conxuntos a risco para cada $x_{(i)}$ por separado e ver así a cantidade de veces que cada dato $x_{(i)}$, $i \in \{1, ..., n\}$, presenta un conxunto a risco unitario. Falemos, por exemplo, dos conxuntos a risco no primeiro dato de cada simulación.

En R xeraranse 1000 valores, os cales corresponden aos resultados dos conxuntos a risco de $x_{(1)}$ nas 1000 simulacións. Esta distribución é moi importante debido a que é a que comeza a xerar a probabilidade de fallo e de supervivencia do primeiro dato, e a que da lugar ao cálculo das probabilidades do resto de datos.

Para $n = 45$, temos que a variabilidade dos conxuntos a risco no primeiro dato de cada mostra é moi grande. Hai casos moi extremos, como poden ser cando aparecen conxuntos a risco moi pequenos, como é o caso da simulación 32, sendo $\frac{1}{3}$ a probabilidade que se leva ese dato, ou incluso na simulación 207, por exemplo, onde se levaría toda a probabilidade. Isto daría lugar a que o resto de datos presentaran unha probabilidade de supervivencia moi baixa ou incluso nula, resultando esto un pouco contradictorio co feito de que os datos foron observados.

Recollemos no Cadro 2.4 un resumo dos conxuntos a risco de $x_{(1)}$ para as 1000 simulacións fixando diferentes tamaños mostrais. A primeira fila indícanos a cantidade de datos que presenta o conxunto a risco de $x_{(1)}$ e a segunda fila as veces que os conxuntos a risco de $x_{(1)}$ teñen ese número de datos. A última columna amósanos a cantidade de veces que $x_{(1)}$ presenta un conxunto a risco con 10 ou máis datos.

Tamaño conxuntos a risco	1	2	3	4	5	6	7	8	9	$\geq$10
Frecuencia para n=45	10	12	32	52	56	62	66	79	85	546
Frecuencia para n=90	2	12	19	19	21	29	30	42	45	781
Frecuencia para n=120	3	12	12	16	28	12	22	31	32	832

Cadro 2.4: Tamaño dos conxuntos a risco para a observación $x_{(1)}$ coas súas respectivas frecuencias para diferentes valores de n.

Observamos entón, á vista do Cadro 2.4, que para $n = 45$ existen 10 conxuntos unitarios en 1000 mostras para o primeiro dato, é dicir, existen 10 mostras de 1000 cun tamaño de $n = 45$, as cales non podemos estimar, posto que recae no dato máis pequeno toda a probabilidade.

Podemos fixarnos na segunda columna do Cadro 2.4. Esta reflíctenos que hai 12 mostras de 1000 cun tamaño tamén de $n = 45$, cuxa primeira observación ten conxunto a risco formado por 2 datos, é dicir, presenta unha probabilidade de $\frac{1}{2}$, sendo esta moi alta e deixando aos 44 datos restantes con probabilidade total de $\frac{1}{2}$ para repartirse.

Observamos tamén no Cadro 2.4 que para $n = 90$ e $n = 120$ xa soamente existen 2 e 3 conxuntos a risco unitarios para $x_{(1)}$, respectivamente, nas 1000 mostras simuladas. Podemos concluír que a existencia de conxuntos a risco cun tamaño pequeno para o primeiro dato da mostra vai diminuindo cando aumentamos o número de datos observados.

Por outra parte, pode resultarnos útil para o estudo crear un código en R, indicado no Apéndice, o cal detecte a cantidade de conxuntos a risco unitarios que existen nas diferentes simulacións, non só no primeiro dato da mostra ordeada, $x_{(1)}$, senón en calquera posición da mostra. Nótese que son os que poden chegar a producir problemas á hora de facer un estudo. Unha vez executado este para $S = 1000$, R devolveranos o número de mostras que presentan algún conxunto a risco unitario, recollidos no Cadro 2.5.

Tamaño da mostra observable (n)	Número de mostras con algún conxunto a risco unitario
45	21
90	6
120	5

Cadro 2.5: Número de mostras que presentan conxuntos a risco unitarios dependendo do n fixado.

A existencia de conxuntos a risco unitarios non suele ser moi frecuente cando se traballa con mostras grandes, debido a que hai máis agrupación de datos provocando así que non se ocasione unha gran cantidade de ocos, como podemos observar no Cadro 2.5. De aquí xorde o Lema 1.4 da sección 1.3, o cal nos indicaba que a medida que aumenta o tamaño mostral, a probabilidade de que haxa conxuntos a risco unitarios vai tender a cero.

2.3. Estimacións do tamaño da poboación

Na seguinte sección imos estimar o tamaño da poboación N a partir da nosa mostra de n datos observados. Para iso o que faremos será fixar N, obtendo valores de n aleatorios que estimarán a $\hat{N}_n$. Para facer estas aproximacións, efectuaremos S simulacións:

$$n_1 \longrightarrow \hat{N}_1$$
$$n_2 \longrightarrow \hat{N}_2$$
$$\vdots$$
$$n_S \longrightarrow \hat{N}_S$$

Obteríamos así S realizacións do noso estimador $\hat{N}_n : \hat{N}_1, \hat{N}_2, ..., \hat{N}_S$. Para calcular o estimador, na sección 1.4 vimos que unha posible opción era a partir do tamaño da mostra n e do estimador $\hat{\alpha}_n$, o cal depende das funcións $\hat{F}_n$ e $\hat{G}_n$. O que faremos entón será crear en R un código, reflexado no Apéndice, o cal nos calcule estes estimadores de F e G, expresados nas fórmulas (1.7) e (1.8), e poder así ter un estimador de α_n, e en consecuencia, un estimador para N.

Podemos atoparnos co problema da aparición de conxuntos a risco unitarios ou pequenos, o cal prexudícanos á hora de calcular o noso estimador $\hat{N}_n$. O que faremos no código para solucionalo será sumarlle unha pequena cantidade, como pode ser $\frac{1}{n}$, aos conxuntos a risco tanto na estimación de F, $\hat{F}_n$, como na de G, $\hat{G}_n$.

Exemplo 2.4. Sexa $(X_1, Y_1), ..., (X_N, Y_N)$ a poboación dada. Consideremos F e G distribucións expoñenciais con parámetros $\lambda_X = \frac{1}{3}$, $\lambda_Y = \frac{1}{3}$, respectivamente. Fixamos o tamaño da poboación, N, e xerando $S = 1000$ simulacións, faremos unha análise do estimador $\hat{N}_n$.

Por un lado, R devolveranos o número de individuos observados, n, para cada simulación, é dicir, aqueles que cumpren a condición $Y \leq X$.

Por outra parte, como dixemos anteriormente, temos que crear un estimador para α_n, o cal recolle a probabilidade de observación estimada en cada unha das simulacións. Podemos calcular analíticamente a expresión de α para a distribución tomada, facendo uso das funcións

$$G(x) = 1 - e^{-\frac{x}{3}}$$
$$dF(x) = f(x)dx = \frac{1}{3}e^{-\frac{x}{3}}dx$$

calculadas no Exemplo 2.1. Teremos así que:

$$P(Y \leq X) = \int_0^\infty G(x)dF(x) = \int_0^\infty (1 - e^{-\frac{x}{3}})\frac{1}{3}e^{-\frac{x}{3}}dx =$$
$$= \int_0^\infty \frac{1}{3}e^{-\frac{x}{3}}dx - \int_0^\infty \frac{1}{3}e^{-\frac{2x}{3}}dx = 1 - \frac{1}{2} = \frac{1}{2}$$

A probabilidade de observación é $\alpha = \frac{1}{2}$, como é obvio, xa que X e Y presentan a mesma distribución. Imos fixar distintos tamaños de poboación, como por exemplo, $N = 100$, $N = 1000$ e $N = 10000$. R devolveranos os seguintes valores para o estimador $\hat{\alpha}_n$.

Para $N = 100$:

$$\hat{\alpha}_n : 0{,}512, 0{,}395, ..., 0{,}595$$

Para $N = 1000$:

$$\hat{\alpha}_n : 0{,}550, 0{,}499, ..., 0{,}359$$

Para $N = 10000$:

$$\hat{\alpha}_n : 0{,}492, 0{,}502, ..., 0{,}512$$

Podemos ver entón, no Cadro 2.6, propiedades deste estimador como poden ser o sesgo, a varianza e o erro cadrático medio, para os distintos valores fixados do tamaño da poboación. Observamos que o sesgo vai diminuíndo a medida que o tamaño da poboación aumenta, e a varianza tende a cero, sendo así $\hat{\alpha}_n$ consistente para a estimación do noso α.

N	$\widehat{sesgo}(\hat{\alpha}_n)$	$\widehat{var}(\hat{\alpha}_n)$	$\widehat{ecm}(\hat{\alpha}_n)$
100	0.0073	0.0220	0.0221
1000	0.0002	0.0030	0.0030
10000	0.0007	0.0006	0.0006

Cadro 2.6: Sesgo, varianza e erro cadrático medio de $\hat{\alpha}_n$ para distintos tamaños de poboación.

Por outro lado, e estimando N da forma $\hat{N}_n = \frac{n}{\hat{\alpha}_n}$, podemos facer unha análise das súas propiedades, para estimar así o tamaño da poboación orixinal. Tomamos por exemplo, o tamaño de poboación $N = 10000$, e o programa R devólvenos un sesgo moi elevado do estimador $\hat{N}_n$ para as 1000 mostras simuladas, como podemos ver no Cadro 2.7. Isto é debido a que existen valores atípicos nas estimacións, xerados pola existencia de conxuntos a risco unitarios, afectando así na estimación de N.

$\widehat{sesgo}(\hat{N}_n)$	$\widehat{var}(\hat{N}_n)$	$\widehat{ecm}(\hat{N}_n)$
51611.41	$2.66{\cdot}10^{12}$	$2.67{\cdot}10^{12}$

Cadro 2.7: Sesgo, varianza e erro cadrático medio de $\hat{N}_n$ para un tamaño de $N = 10000$.

O que faremos entón será buscar estas estimacións que se presentan moi alonxadas do noso N fixado e eliminarémolas. Para iso, a través de códigos en R, buscaremos as estimacións máis grandes de $\hat{N}_n$ e as máis pequenas, tendo así os seguintes resultados:

```
# Estimacións máis pequenas
[1] 9058.723    9178.342    9222.174    9230.109    9243.961    9251.370
[7] 9253.039    9287.957    9290.763    9294.632
# Estimacións máis grandes
[1] 11169.49    11261.32    11286.00    11292.76    11339.86    11350.73
[7] 11639.95    11657.71    12017.96    12665.68    51632414.10
```

Observamos que o último dos valores das "estimacións máis grandes" é moi alto e moi distinto ao noso valor fixado da poboación, que era $N = 10000$. Logo, o que faremos nestas situacións será fixar unha condición, na cal se eliminarán todas aquelas estimacións de N que sexan superiores a 10 veces o valor de n, xa que se considerarán atípicas, e así poderemos obter un mellor estimador $\hat{N}_n$. Para $N = 10000$ temos que soamente habería que eliminar o dato 51632414.10, xa que é o único que cumpre a condición. Imos facer isto entón para os distintos tamaños de poboación fixados anteriormente e ver a cantidade de estimacións que teríamos que quitar en cada un dos casos. Calcularemos tamén o sesgo, a varianza e o erro cadrático medio do estimador $\hat{N}_n$ para ver se eliminando os datos atípicos, a aproximación do tamaño de poboacion sería mellorada. Podemos ver esto reflexado no Cadro 2.8.

Tamaño fixado da poboación (N)	Cantidade de estimacións que temos que eliminar	$\widehat{sesgo}(\hat{N}_n)$	$\widehat{var}(\hat{N}_n)$	$\widehat{ecm}(\hat{N}_n)$
100	22	4.819	1561.465	1584.692
1000	1	12.995	19145.08	19313.95
10000	1	-11.014	143189.7	143311

Cadro 2.8: Sesgo, varianza e erro cadrático medio do noso estimador $\hat{N}_n$ e cantidade de estimacións que temos que eliminar porque cumpren a condición fixada de ser atípicas.

Anteriormente, no Cadro 2.7, tiñamos para $N = 10000$ un sesgo moi elevado debido ás estimacións atípicas, e podemos observar, no Cadro 2.8, que quitando o dato que amosaba unha mala estimación de N, o sesgo diminuiu considerablemente, obtendo así unha milloría do estimador do tamaño da poboación. Vemos que a varianza tamén diminúe en comparación co Cadro 2.7. No Cadro 2.8 a varianza vai aumentando a medida que aumenta o que queremos estimar, é dicir, N.

Á vista do Cadro 2.8, vemos que o aumento do tamaño de poboación fai que nos atopemos con unha cantidade menor de valores atípicos. Unha vez eliminados estes, para cada un dos N tamaños fixados, podemos obter a esperanza e a varianza de $\frac{\hat{N}_n}{N}$, como reflexa o Cadro 2.9.

Tamaño fixado da poboación (N)	$\mathbb{E}(\frac{\hat{N}_n}{N})$	$\widehat{var}(\frac{\hat{N}_n}{N})$
100	1.034	0.117
1000	1.013	0.019
10000	0.999	0.001

Cadro 2.9: Media e varianza de $\frac{\hat{N}_n}{N}$.

Podemos observar no Cadro 2.9 que a medida que o tamaño da poboación aumenta, temos que a esperanza $\frac{\hat{N}_n}{N}$ vai estar en torno a 1, como reflexaba o Corolario 1.8 da sección 1.5. Tamén vemos que a varianza diminúe cando N aumenta, tendendo esta a cero.

Podemos concluír que o estimador $\hat{N}_n$ nos dará unha boa estimación do tamaño da nosa poboación orixinal unha vez elimados aqueles valores que consideremos como atípicos.

Observación 2.5. No Exemplo 2.4 traballamos cunha distribución expoñencial con parámetro λ iguais para ambas distribucións de X e Y. Supoñamos F e G con parámetros $\lambda_X = \frac{1}{2}$ e $\lambda_Y = \frac{1}{4}$, respectivamente.

Neste caso a probabilidade de observación non vai ser a mesma que no Exemplo 2.4, xa que X e Y non presentan a mesma distribución. Temos entón que o valor de α virá dado por:

$$P(Y \leq X) = \int_0^\infty G(x)dF(x) = \int_0^\infty (1 - e^{-\frac{x}{4}})\frac{1}{2}e^{-\frac{x}{2}}dx =$$
$$= \int_0^\infty \frac{1}{2}e^{-\frac{x}{2}} - \int_0^\infty \frac{1}{2}e^{-\frac{3x}{4}}dx = 1 - \frac{4}{6} = \frac{1}{3}$$

Vemos que a probabilidade de observación é menor que no caso anterior, sendo $\alpha = \frac{1}{3}$. Polo tanto teremos unha menor cantidade de datos observados e en consecuencia, un maior truncamento da poboación orixinal.

Podemos realizar un estudo similar ao do Exemplo 2.4, fixando N. Tomaremos os tamaños de poboación $N = 100$, $N = 1000$ e $N = 10000$ e fixaremos a mesma condición que antes para as aproximacións atípicas, eliminando aquelas que sexan superiores a 10 veces o valor de n, é dicir, que sexan 10 veces máis grandes que os datos observables.

Neste caso, a cantidade de valores atípicos aumentaron considerablemente para os diferentes N con respecto ao Exemplo 2.4, como resulta lóxico, xa que agora a probabilidade de observación é menor, tendo menos datos observados e perxudicando isto á hora de estimar o tamaño da poboación orixinal.

Podemos ver reflexados no Cadro 2.10 a esperanza e varianza de $\frac{\hat{N}_n}{N}$, e facer así unha comparación co Cadro 2.9.

Tamaño fixado da poboación (N)	$\mathbb{E}(\frac{\hat{N}_n}{N})$	$\widehat{var}(\frac{\hat{N}_n}{N})$
100	1.014	0.323
1000	1.002	0.074
10000	0.995	0.016

Cadro 2.10: Media e varianza de $\frac{\hat{N}_n}{N}$.

Observamos, ao igual que pasaba no Cadro 2.9, que a esperanza de $\frac{\hat{N}_n}{N}$ mantense ao redor de 1. Podemos fixarnos tamén na terceira columna do Cadro 2.10, onde a varianza tende a cero a medida que N aumenta. Podemos comparar esta varianza ca do Cadro 2.9, e vemos que é maior nesta táboa para cada un dos N fixados, estando isto relacionado coa diminución da probabilidade de observación dos datos con respecto ao exemplo anterior. Polo tanto, podemos concluír que a probabilidade de observación afecta á hora de atopar un bon estimador para estimar o tamaño total da poboación.

Capítulo 3

APÉNDICE

3.1. Función de supervivencia, función de risco e función de risco acumulado

A función de supervivencia defínese como a probabilidade de que un individuo sobreviva máis alá dun instante x, é dicir:

$$S(x) = P(X > x)$$

Sabemos que a función de distribución se define como $F(x) = P(X \leq x)$ e representa, neste caso, a probabilidade de que un individuo non sobreviva a ese instante x. Logo podemos relacionar a función de distribución coa función de supervivencia da seguinte forma:

$$F(x) = 1 - S(x)$$

Cando X é unha variable aleatoria discreta e toma valores $x_1 < x_2 < ...$, podemos calcular a función de supervivencia e a súa correspondente función de distribución como a suma de funcións de masa de probabilidade $p(x_i) = P(X = x_i)$ da seguinte maneira:

$$S(x) = P(X > x) = \sum_{x_i > x} p(x_i) \qquad F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

No caso discreto, a función de risco, tamén chamada razón de fallo, defínese como a probabilidade de caer no instante x_i sabendo que se chegou ata ese tempo. Podemos relacionala coa supervivencia do seguinte xeito:

$$\lambda_F(x_i) = P(X = x_i / X \geq x_i) = \frac{P(X = x_i, X \geq x_i)}{P(X \geq x_i)} = \frac{P(X = x_i)}{P(X > x_{i-1})} =$$

$$= \frac{P(X > x_{i-1}) - P(X > x_i)}{P(X > x_{i-1})} = \frac{S(x_{i-1}) - S(x_i)}{S(x_{i-1})} = 1 - \frac{S(x_i)}{S(x_{i-1})} \qquad i \in \{1, 2, ..., n\}$$

A expresión anterior permite obter de maneira recursiva a función de supervivencia en funcións dos valores da razón de fallo, pois $S(x_0) = 1$, $S(x_1) = 1 - \lambda(x_1)$, $S(x_2) = S(x_1)(1-\lambda(x_2))$,..., chegando así a que a supervivencia podémola expresar como producto de funcións de risco:

$$S(x) = \prod_{x_i \leq x} (1 - \lambda(x_i)), \qquad x \in \mathbb{R} \tag{3.1}$$

Ademais, podemos definir a función de risco acumulativo no caso discreto como:

$$\Lambda(x) = \sum_{x_i \leq x} \lambda(x_i) \tag{3.2}$$

Por outra banda, cando X é unha variable aleatoria continua con densidade $f(x)$, temos que a función de supervivencia e a función de distribución son:

$$S(x) = P(X > x) = \int_x^{\infty} f(x)dx \qquad F(x) = P(X \leq x) = \int_0^x f(x)dx$$

Neste caso temos que a función de risco se define como o cociente entre a función de densidade e a función de supervivencia, tendo así que:

$$\lambda_F(x) = P(X = x_i / X \geq x_i) = \frac{f(x)}{P(X \geq x)} = \frac{f(x)}{1 - F(x^-)}$$

No caso de que X sexa absolutamente continua con densidade f, a función de risco acumulativo vén dada por:

$$\Lambda(x) = \int_0^x \lambda(z)\mathrm{d}z$$

Como $\lambda(x) = \frac{f(x)}{1-F(x^-)}$, para 0 $\leq$ x $<\infty$, obtemos:

$$\Lambda(x) = \int_0^x \frac{f(z)}{1-F(z)}\mathrm{d}z = -\int_0^x \frac{-f(z)}{1-F(z)}\mathrm{d}z = |-log(1-F(z))|_0^x =$$
$$= -log(1-F(x)) + log(1-F(0)) = -log(1-F(x))$$

xa que F(0)=0.

Tomando expoñenciais:

$$-\Lambda(x) = log(1-F(x)) \Rightarrow exp(-\Lambda(x)) = 1 - F(x)$$

obtendo así que:

$$S(x) = exp(-\Lambda(x))$$

No caso onde a variable teña parte discreta e parte continua, a función de supervivencia pódese expresar como:

$$1 - F(x) = \underbrace{\prod(1 - \lambda(z))}_{(a)} \underbrace{exp(-\Lambda_c(x))}_{(b)}$$

sendo (a) a función de risco instantáneo e (b) a función de risco continua. Concluímos así que a función de risco caracteriza a distribución, e deducimos a maneira de obter a función de supervivencia ou de distribución a partir da función de risco.

Podemos distinguir varios casos dependendo do valor da función de fallo:

1. Fallo inicial: manifféstase ao inicio e a medida que avanza o tempo, desaparece.

2. Fallo accidental: función de risco constante a medida que pasa o tempo e suele ser menor que a que presentaba ao inicio.

3. Fallo de desgaste: a función de risco vai aumentando co paso do tempo.

Exemplo 3.1. Ilustremos esto coa distribución expoñencial, a cal é unha distribución de probabilidade continua con parámetro $\lambda > 0$ e función de densidade $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. A súa función de distribución vén dada por:

$$F(x) = \int_0^x \lambda e^{-\lambda x} = 1 - e^{-\lambda x}$$

e polo tanto, a súa función supervivencia será:

$$S(x) = e^{-\lambda x}$$

Sabemos, como vimos anteriormente, que a función de risco da distribución expoñencial será o cociente entre $f(x)$ e $S(x)$, neste caso:

$$\lambda_F(x) = \lambda$$

Cando λ aumente, a función de risco será maior, é dicir, existirá unha maior probabilidade de caer no instante de tempo x sabendo que se chegou a el; aínda que a medida que λ se fai maior, existe unha estabilidade na función de risco respecto a ese λ fixado, como se pode observar na Figura 3.1.

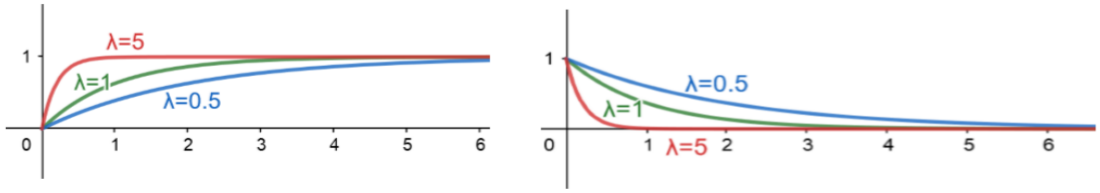


Figura 3.1: Á esquerda a función de distribución da expoñencial e á dereita a función de supervivencia da expoñencial.

3.2. Función de risco en tempo invertido

Podemos falar tamén da función de risco en tempo invertido, que é a probabilidade de caer no instante y_i invertindo o tempo, é dicir, tomando o tempo dende un punto final cara atrás.

No caso discreto vén dada por:

$$\lambda_G^i(y_i) = P(Y = y_i / Y \leq y_i) = \frac{P(Y = y_i, Y \leq y_i)}{P(Y \leq y_i)} = \frac{P(Y = y_i)}{P(Y \leq y_i)} =$$

$$= \frac{P(Y \leq y_i) - P(Y \leq y_{i-1})}{1 - P(Y > y_i)} = \frac{G(y_i) - G(y_{i-1})}{G(y_i)} = 1 - \frac{G(y_{i-1})}{G(y_i)}$$

chegando así a que:

$$G(y) = \prod_{y_i > y} (1 - \lambda_G^i(y_i)), \qquad y \in \mathbb{R}$$

mentras que no caso continuo será:

$$\lambda_G^i(y) = P(Y = y_i / Y \leq y_i) = \frac{g(y)}{P(Y \leq y)} = \frac{g(y)}{G(y)}$$

3.3. Códigos de R

Na seguinte sección do apéndice imos amosar as liñas de código R que fumos empregando ao longo do traballo para facer as diferentes simulacións.

Faremos fincapé no código principal empregado para a maioría dos exemplos, onde obtemos resultados sobre distintas propiedades que presentan os estimadores da media de X propostos no capítulo 2 deste traballo.

```
set.seed(123456)

# Distribución expoñencial
lambdax=1/3
lambday=1/3
mediax=1/lambdax

N=60 # Número de datos da poboación orixinal
ns=100  # Número de mostras simuladas

media=c()
media_mala=c()

# Bucle das mostras
for (is in 1:ns){

x=rexp(N,rate=lambdax)
y=rexp(N,rate=lambday)

io=0
xo=c()
yo=c()
for (i in 1:N){
if (y[i]<=x[i]){
io=io+1
yo[io]=y[i]
xo[io]=x[i]
}
}

n=io
n # número de mostras despois do truncamento

plot(xo,yo,xlim=c(0,max(xo)),ylim=c(0,max(xo)))
abline(a=0,b=1)
```

```
# Imos estimar a distribución da X coa mostra das (xo,yo)
z=c(xo,yo)
cn=c()
indice=sort(z,index.return=TRUE)$ix
ix=1
cn[1]=0
xord=c()  # x ordenadas
for (i in 1:(2*n)){
if (indice[i]>n){
cn[ix]=cn[ix]+1
}else{
xord[ix]=z[indice[i]]
lines(c(xord[ix],xord[ix],max(xo)),c(0,xord[ix],xord[ix]),lty=2,col=ix)
ix=ix+1
cn[ix]=cn[ix-1]-1
}
}
cn=cn[1:n]
cn # Conxuntos a risco

# Supervivencias de cada dato do conxunto a risco
superviv=cumprod(1-1/cn)
superviv
F=1-superviv

media_mala[is]=mean(xo)
media_mala[is]
prob=c(1,superviv[1:(n-1)])-superviv
prob
media[is]=sum(xord*prob)
media[is]
}   # Aquí pecha o bucle das mostras simuladas

#Estimacións de mu_T
sesgo_media=mean(media)-mediax; sesgo_media
var_media=var(media); var_media
```

```
ecm_media=var_media+(sesgo_media)^2; ecm_media


#Estimacións de mu_o
sesgo_media_mala=mean(media_mala)-mediax; sesgo_media_mala
var_media_mala=var(media_mala); var_media_mala
ecm_media_mala=var_media_mala+(sesgo_media_mala)^2; ecm_media_mala
```

Podemos observar que o código anterior foi creado para unha distribución expoñencial. Simplemente cunha pequena modificación, podemos facer uso deste coa distribución uniforme no intervalo [0,1]. Basta reemplazar as liñas da expoñencial por:

```
# Distribución uniforme
a=0
b=1
mediax=(a+b)/2


x=runif(N,min=a,max=b)
y=runif(N,min=a,max=b)
```

Neste código o que fixemos foi fixar N e ver como variaba o número de mostras observables, n, despois de aplicar o truncamento, sacando así resultados de propiedades como o sesgo, varianza e erro cadrático medio de ambos estimadores da media, μ_o e μ_T.

Deseguido imos ilustrar as liñas que hai que variar na bucle do código para cando o que se mantén fixo é o número de datos observados, n, coa distribución expoñencial, obtendo así os resultados do Exemplo 2.1.

```
for (is in 1:ns){
io=0
xo=c()
yo=c()
while(io<n){  x=rexp(1,rate=lambdax)
  y=rexp(1,rate=lambday)
if (y<=x){
io=io+1
yo[io]=y
xo[io]=x
}
}
}
```

Como vemos consistiría en cambiar o comando *for* dentro do bucle por *while*.

O cambio que se produce así é que no primeiro dos casos, o bucle comezaría co comando *for* repetindo a acción de xerar 100 veces, no caso do Exemplo 2.1, N mostras con distribución expoñencial e logo tomaríanse aquelas que cumpliran a condición de truncamento, obtendo así n. Se cambiamos isto por *while*, o que fará o noso programa será que mentras que se teña un número inferior ao das mostras observadas, $io < n$, xeraranse estas con distribución expoñencial ata que se deixe de cumprir $io < n$, e logo verase que verifican a condición de truncamento. En ambos casos, finalizarase o bucle sendo $io = n$.

Falaremos, a continuación, das liñas de código empregadas no caso das estimación dos conxuntos a risco unitarios, que se corresponden coa sección 2.2 do segundo capítulo do traballo. O que se fixo foi crear unha matriz, *matcn*, na cal se almacenan os conxuntos a risco de cada dato en cada unha das mostras simuladas, sendo as filas o número de mostras xeradas e as columnas os individuos da mostra ordeada. Logo creouse un contador, *detcn1*, o cal nos devolve o resultado da suma de conxuntos a risco unitarios que hai en todas as simulacións, estando fixado o valor de n. Vexamos entón o código de R empregado:

```
matcn=matrix(0,nrow=ns,ncol=n)
detcn1=0
cn # Conxuntos a risco

ncn1=sum(cn==1)  # Número total de conxuntos a risco unitarios
if (ncn1>1){detcn1=detcn1+1}  # Detectamos un cn unitario non trivial
```

Para a última sección das simulacións, para o cálculo de α_n, foi necesario o cálculo de G, polo tanto no código principal añadíronse os mesmos cálculos que para F pero tendo en conta que os conxuntos a risco neste caso están en tempo invertido. Unha vez obtidas as estimacións das distribucións de X e Y, estimamos α_n, e en consecuencia, o tamaño da poboación, N.

```
iy=0
ix=0
alphan=0
for (i in 1:(2*n)){
if (indice[i]>n){
iy=iy+1
Gx=G[iy]
}else{
```

```
ix=ix+1
alphan=alphan+Gx*prob[ix]
}
}
valphan[is]=alphan
Nhat[is]=n/alphan
```

Unha vez programado todo isto, pódense calcular todas as propiedades que desexemos destes estimadores.

Bibliografía

[1] Woodroofe, M., *Estimating a distribution function with truncated data*, The Annals of Statistics, Vol.13 (1985), 163–177.

[2] A. Taboada, X., Fernández, E., González Manteiga, W., Hervada, X., L. Otero, X., Sánchez Sellero, C., Vázquez Fernández, E., *Reporting delay: A review with a simulation study and application to spanish aids data*, Statistics in Medicine, Vol.15 (1996), 305–321.

[3] Strzalkowska-Kominiak, E., Stute, W., *On the probability of holes in truncated samples*, Journal of Statistical Planning and Inference 140 (2010), 1519–1528.

[4] Lynden-Bell, D., *A method of allowing for known observational selection in small samples applied to 3cr quasars*, Mon. Not. R. astr. Soc., I55 (1971), 95–118.